THE OFFICIAL MAGAZINE OF THE OCEANOGRAPHY SOCIETY

Oceanography

CITATION

Sedberry, G.R., D.G. Fautin, M. Feldman, M.D. Fornwall, P. Goldstein, and R.P. Guralnick. 2011. OBIS-USA: A data-sharing legacy of the Census of Marine Life. *Oceanography* 24(2):166–173, doi:10.5670/oceanog.2011.36.

BY GEORGE R. SEDBERRY, DAPHNE G. FAUTIN, MICHAEL FELDMAN, MARK D. FORNWALL, PHILIP GOLDSTEIN, AND ROBERT P. GURALNICK

# OBIS-USA

## A Data-Sharing Legacy of the Census of Marine Life

A blackbelly rosefish, *Helicolenus dactylopterus* (Delaroche, 1809), photographed using a remotely operated vehicle at a depth of 197 m off the coast of South Carolina (32°43.8'N, 78°05.5'W). Figure 1 shows the distribution of blackbelly rosefish, as determined from fishery-monitoring trawl surveys. *Photo credit: NOAA*

ABSTRACT. The United States Geological Survey's Biological Informatics Program hosts OBIS-USA, the US node of the Ocean Biogeographic Information System (OBIS). OBIS-USA gathers, coordinates, applies standard formats to, and makes widely available data on biological collections in marine waters of the United States and other areas where US investigators have collected data and, in some instances, specimens. OBIS-USA delivers its data to OBIS international, which then delivers its data to the Global Biodiversity Information Facility (GBIF) and other Web portals for marine biodiversity data. OBIS-USA currently has 145 data sets from 36 participants, representing over 6.5 million occurrence records of over 83,000 taxa from more than 888,000 locations. OBIS-USA, a legacy of the decade-long (2001–2010) international collaborative Census of Marine Life enterprise, continues to add data, including those from ongoing Census projects. Among the many challenges in creating OBIS, including OBIS-USA, were developing a community of trust and shared value among data providers, and demonstrating to providers the value of making their data accessible to others. Challenges also posed by the diversity of data sets relevant to marine biodiversity stored on thousands of computers, in a variety of formats, not all widely accessible, have been met in OBIS-USA by implementing a uniform standard and publishing platform that is easily accessible to a broad range of users.

## INTRODUCTION

OBIS-USA (http://obisusa.nbii.gov) is the United States node of the Ocean Biogeographic Information System (OBIS; http://www.iobis.org). Established in 2000 as a project to house data collected by the Census of Marine Life (hereafter called Census; http://www.coml.org), OBIS is an online database that amalgamates and integrates specimen-level data concerning the distribution of marine life throughout the world (Grassle, 2000). It comprises a centralized coordinating office and many regional and thematic nodes (http://www.iobis.org/obis/regional-nodes). It was planned from the outset that OBIS would persist as a legacy of the Census (Yarincik and O'Dor, 2005). As of the end of 2010 and the conclusion of the Census enterprise, OBIS served 30.3 million occurrence records for 120,000 species from 898 data sets (http://www.iobis.org/node/307), which it delivers to the Global Biodiversity Information Facility (GBIF), an open-access database of terrestrial and aquatic biodiversity (http://www.gbif.org), founded about the same time as OBIS. OBIS provides data and other support to partnering Web portals for marine biodiversity data, such as the World Register of Marine Species (WoRMS; http://www.marinespecies.org/about.php) and the Encyclopedia of Life (EOL; http://www.eol.org/content/partners?page=4). The community-based approach used by OBIS is similar to that employed subsequently by the vertebrate database VertNet (http://vertnet.org/index.php) and its component networks that amalgamate museum collection data, but OBIS includes many marine databases in addition to those associated with natural history museum collections.

Recognizing the advantages of a decentralized structure for both acquiring and delivering data (Fornwall, 2000), regional, national, and thematic OBIS nodes were established to aid in finding and gathering data, as well as for communication and coordination. Individual node managers are responsible for building relationships with institutions and individuals in the geographical or thematic purview of that node. An individually built access portal allows each node to offer a data interface to meet its community's needs, including use of appropriate language. Other advantages of a decentralized structure include enhancing speed of data delivery and providing financial stability for the enterprise.

OBIS-USA is part of the United States Geological Survey's Biological Informatics Program, which provides access to information about the nation's living resources, a critical national asset, according to the National Research Council (NRC). NRC recognized in *A Biological Survey for the Nation* (National Research Council, 1993) that much of the considerable amount known about living resources in the United States was not readily accessible; state and local governments, educational and nongovernmental organizations, and the private sector held a significant portion of the data because the federal government had not collected it. The report urged the federal government to play a lead role in promoting the sharing of the data to support natural resources management.

Since its beginning in 2005, OBIS-USA has become the major

provider of US marine data to OBIS. As of January 2011, OBIS-USA served data from 145 data sets contributed by 36 participants, representing over 6.5 million occurrence records of more than 83,000 taxa from over 888,000 locations. Data continue to be added, including those from ongoing Census projects.

In this article, we review challenges articulated by some of the original OBIS participants, then describe the development approach taken by OBIS-USA, and consider future challenges. This review is less to enumerate the accomplishments of OBIS-USA than to impart lessons learned about the importance of sharing data across agencies and projects, and how to bridge community needs. In particular, we focus on the value of encouraging community acceptance of data standards that are sufficiently flexible and extensible to capture the needs of data producers and consumers as their needs change. For consumers, we focus on the value of summary views of data and metadata within and across shared, interoperable data sets to allow quick data discovery, enhancement, and use; for providers, we focus on what OBIS-USA can provide to enhance quality of the data and metadata and make those enhanced data more easily discoverable and interoperable.

## CHALLENGES

In 2000, *Oceanography* published a special issue (Volume 13, Number 3) entitled *Ocean Biogeographic Information System* that provided a glimpse into the state of marine biodiversity informatics at the turn of the twenty-first century. In the lead paper, Grassle (2000) put forth the vision for OBIS as an online, user-friendly system for absorbing, integrating, and assessing data about marine organisms. Authors of successive papers discussed some ongoing marine biological data efforts and the challenges for technological capability, standards development (including taxonomy and geography), data quality, and interoperability of various forms of data. Many authors recognized that a critical step to resolving these challenges was centered on improved and increased engagement of the community of data users and providers.

Sociological challenges can be more problematic than technical ones for those who develop distributed electronic data systems such as OBIS and OBIS-USA. Indeed, Grassle (2000, p. 7) concluded: "The biggest challenge is for the various stakeholders worldwide in marine biological data to learn quickly to work together." A potential participant may be reluctant to share data, fearing they will be published before the scientist who gathered them can do so, or that the data will be misused or misinterpreted, perhaps applied in a context other than that for which they were collected, or subjected to exposure of gaps in coverage or quality. Another source of reluctance to participate is the effort involved in creating appropriate metadata and configuring data to meet the standards or structure of the distributed system. The OBIS-USA process, elaborated below in Solutions and Process, promotes working closely with those who may be initially skeptical about centralizing data for access in order to change such perceptions by showing that an open system promotes maximal return on investment for the high initial costs of collecting and codifying the data.

OBIS-USA obtains data from providers who have amassed the data for a variety of purposes, which creates a set of data management and standardization challenges to be met. For example, some of the extensive biological data from surveys, such as those for fisheries monitoring, are obtained by visual census or through catches discarded at sea once standard information is recorded. Such data may contain taxonomic errors because of the unfamiliarity of fishery surveyors with current literature, or because of taxonomic changes published since the surveys were conducted (e.g., Collette and Vecchione, 1995). Because museum data are based on voucher specimens (which may include specimens from surveys), an identification can be corrected, sex can be determined, and (for many) a DNA sequence can be obtained long after the specimen is collected. However, many museum specimens lack details contained in survey data, such as sampling methods,

**George R. Sedberry** *(george.sedberry@noaa.gov) is Superintendent, Gray's Reef National Marine Sanctuary, National Oceanic and Atmospheric Administration, Savannah, GA, USA.* **Daphne G. Fautin** *is Professor, Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, USA.* **Michael Feldman** *is Program Manager, Census of Marine Life, Consortium for Ocean Leadership, Washington, DC, USA.* **Mark D. Fornwall** *is Senior Advisor, Biological Informatics Program, United States Geological Survey, Kahului, HI, USA.* **Philip Goldstein** *is Informatics Manager, University of Colorado at Boulder, Boulder, CO, USA.* **Robert P. Guralnick** *is Curator of Invertebrate Zoology, University of Colorado Museum of Natural History, Boulder, CO, USA.*

abundance, and precise collection locality. Integrating data from diverse sources requires in-depth understanding of the complex differences among data sets and broadening metadata standards to include both the simplistic and the very rich data delivered by providers. Understanding and managing for these differences is essential for creating an integrated, useful, and well-defined data system.

A related scientific challenge, discussed by several authors in the 2000 *Oceanography* volume, is taxonomic and nomenclature change and uncertainty. OBIS-USA receives data at various levels of taxonomic detail. Some participants identify a full taxonomic hierarchy; some provide only scientific name. Data collectors and providers rely on many taxonomic authorities and methods to make identifications, but information on how the identification was made is often not included with the specimen record. Despite the best efforts of the data providers, taxonomic detail may contain what appear to be errors, such as multiple spellings for a taxon, or using out-of-date nomenclature. These errors may decrease the value and usefulness of the data sets. OBIS-USA staff screens data upon receipt, and alerts participants to possible taxonomic issues. In most cases, this review does not result in any immediate change to data by the participant, but it may generate valuable follow-up tasks for the future that participants can prioritize and conduct as suits them. OBIS-USA is currently developing approaches to check taxonomic names as provided by participants against the Integrated Taxonomic Information System (ITIS; http://itis.gov), as a reference. ITIS is a partnership of agencies and organizations such as the Catalogue of Life (COL; http://www.catalogueoflife.org) from the United States, Canada, and Mexico, with experts from around the world, that has created a database of scientific and common names of biota of interest to North America. ITIS, like OBIS-USA, is part of the US National Biological Information Infrastructure (http://www.nbii.gov) and an associate member of GBIF. OBIS-USA staff consults with these additional biodiversity databases for taxonomic consistency. The goal of this check is not to change original data, but to determine if cross-referencing to ITIS can enhance search and usability of data in OBIS-USA.

A final challenge, one that is critical to engaging users, is timely response to queries. In the early stages of OBIS, the system was operationally distributed, so each query received was sent to all data providers. As the number of providers increased, response time slowed markedly. Additionally, if one provider did not respond, the entire query failed. Even for successful queries, the extent of the network successfully accessed was unknowable, and metrics of data download and use could not be tracked (Constable et al., 2010).

## SOLUTIONS AND PROCESS

OBIS-USA has overcome technical and sociological challenges to develop a functioning ocean biogeographic data-sharing community that engages a wide range of participants. Collaboration is encouraged through clear explanations to potential data providers and users of the many benefits of participation in the OBIS-USA network. Recruitment of participants who are new to the concept of data sharing has required a proactive process that ensures members of the community are included in each step of developing OBIS-USA. Careful listening by the staff allows participants' needs to be addressed by adapting protocols to handle issues such as those concerning data standardization.

Foremost among the numerous benefits to both users and providers of integrated data is identification of errors and areas of concern; this evaluation often occurs through the opportunity to review data with a set of fresh eyes. Formatting procedures help to spot, for example, geographic outliers and taxonomic inconsistencies. Inconsistencies are reported to the data provider, who can then examine the data set for possible errors. A report of taxonomic inconsistency, for example, seldom results in an immediate change, but, as time becomes available, the participant may follow up and subsequently provide cleaner data sets to OBIS-USA. By engaging in this process, participating institutions become more aware of external data standards and practices that might be useful in collecting and processing data. As trust and understanding increase between OBIS-USA and data providers, long-term relationships and feedback loops develop so providers improve their data collection and archiving practices. In turn, subsequent data transfer to OBIS-USA becomes more efficient and thus helps encourage additional participation from existing and new data providers. Tools being developed will automate some aspects of dealing with taxonomic and nomenclatural issues, thereby eliminating the sometimes-significant costs of human intervention to alter records. Better documentation of how data are systematically collected

using standard vocabularies has minimized taxonomic problems: common names used in commercial trade represent a standard taxonomy that is both highly reproducible and easily translatable to a scientific name.

Engagement with data providers and users fulfills two critical elements of OBIS-USA: creating metadata to help users discover the data and assess its suitability for use, and assisting providers in making available to the public data that are of the highest quality possible. Serving high-quality data can involve costs (including effort) to the provider, although these costs are often overestimated. OBIS-USA offers a user-friendly interface, and devotes significant staff time to assisting providers with metadata preparation and mapping data to the OBIS data schema (http://www.iobis.org/data/schema-and-metadata) so that data are fully captured while minimizing the burden on providers. Most providers are willing to improve data. The result is community access to a set of data that is spatially, taxonomically, and temporally as comprehensive and error-free as possible. Data gaps revealed through the collaboration provide a basis for future research. In addition, participants join a community of skilled data collectors and processors, which provides an opportunity for sharing of skills and knowledge.

Recognizing the primary importance of licensing, data attribution, and citation, OBIS-USA works with its providers to establish proper intellectual property practices. Attributions and data-use constraints are included for each data set by the contributor, and these are incorporated into metadata that are compliant with Federal Geographic Data Committee standards. A user who requests data agrees to constraints and proper citation requirements upon clicking the download button. OBIS-USA will not make public any participant's data until these elements of metadata are in place; this is the most rigid requirement OBIS-USA places on metadata. With regard to use and citation agreements for data delivered via Web services (see below), OBIS-USA includes the participant's identification, citation, and use limitations in every record returned by the service. OBIS-USA is evaluating whether such citation information is excessive, but initial results suggest that providing it is prudent.

## USING OBIS-USA

The OBIS-USA Web site allows users to search and acquire data sets or individual records of interest. Current access tools are effective, but OBIS-USA continues to introduce new search tools that accomplish much more: some are already available as beta test on the public site. Through these search tools, an OBIS-USA user can select data sets by spatial criteria and search on fields such as scientific name (Figure 1). OBIS-USA companion tools summarize characteristics of every data set, sorted by the attributes, quantity, and quality of the contents, down to individual fields. Users may review quality by using applications that provide presentations of taxonomic depth and duplication of data for each data set. Tools allow for some simple analyses of the data concerning spatial extent (in three dimensions) and taxonomic classification. While not eliminating problems concerned with taxonomic and geospatial precision, this approach allows a user to assess each data set or record for its fitness of use for his/her needs.

A user can evaluate and download records that lie within geographical boundaries drawn using a familiar Internet mapping interface. A quick analysis and mapping option informs the user about the number of records in that geographic area and the data sets from which they come (Figure 2). The user will have the option of downloading the entire selection or individual data sets. More in-depth analysis, which can take a few seconds or a few minutes depending on the scope of the search, enables the user to see and download data within the geographic bounding box by data set, date, depth, and taxon (full taxon list, or the 25 most frequent, or predominant higher-taxon groups).

When a user obtains results of searches by taxon or geographic area, sorted by the data sets in which records are found, s/he has three options to download and/or explore the data further. The user can (1) download all the records that match search criteria for that data set; (2) view metadata that describe purpose, methods, and other details about the data set relevant to both discovery and use (Figure 2); and (3) refer to additional Web-based resources or publications related to the OBIS-USA representation of the data set, including the collection of original research data products (e.g., reports, publications), and hydrographic measurements (e.g., depth, temperature, salinity), if available. The third option is still under development. Enthusiasm for this function by participants, some of whom have offered to provide such links for their own related resources, has helped make tangible the community benefit that is the goal of such an offering.

It is critical to the usability of OBIS-USA that all the features for downloading data are enabled for Web services. The primary objective of Web service access is to allow machine-to-machine access so a user can automate mapping or analysis, and create searches by criteria other than those available through the OBIS-USA site. Web services also provide a means of integrating biological data served by OBIS-USA with hydrographic, meteorological, human-use, and other such data. OBIS-USA makes it possible for biological information and multiple other data types (e.g., temperature or salinity) that share spatial and temporal data (e.g., coordinates or date) to be queried simultaneously, an essential step in coastal and marine spatial planning. Indeed, previous investigations using the OBIS database show that sea surface temperature is an important environmental predictor of marine biodiversity (Tittensor et al., 2010). This approach needs further investigation on a finer scale, and in relation to other environmental variables as those data become available at OBIS-USA.

The initial Web service implementation for OBIS-USA uses the National Oceanic and Atmospheric Administration (NOAA) Environmental Research Division's Data Access Program (ERDDAP) Web service facility (http://coastwatch.pfeg.noaa.gov/erddap/index.html). The OBIS-USA ERDDAP services are at http://obis-usa.colorado.edu/erddap/index.html. ERDDAP allows access to multiple data types and their integration using standard and commonly used formats, including (but not limited to) Network Common Data Form (NetCDF), MATLAB, and Keyhole Markup Language (KML). Once in a common format, different types of data can interact in integrated models and applications.

## THE FUTURE OF OBIS-USA

OBIS-USA has a thriving and growing data-sharing community that is heterogeneous in terms of sources, including natural history museums, fisheries agencies, ocean data centers, and the private sector. It thus represents a unique resource for those interested in ocean biodiversity and natural resource management.

The challenges related to data sharing and quality are slowly being addressed, but much remains be done. Two key areas that need continued attention are integrating geographic and environmental data with the biological data more fully, and developing more

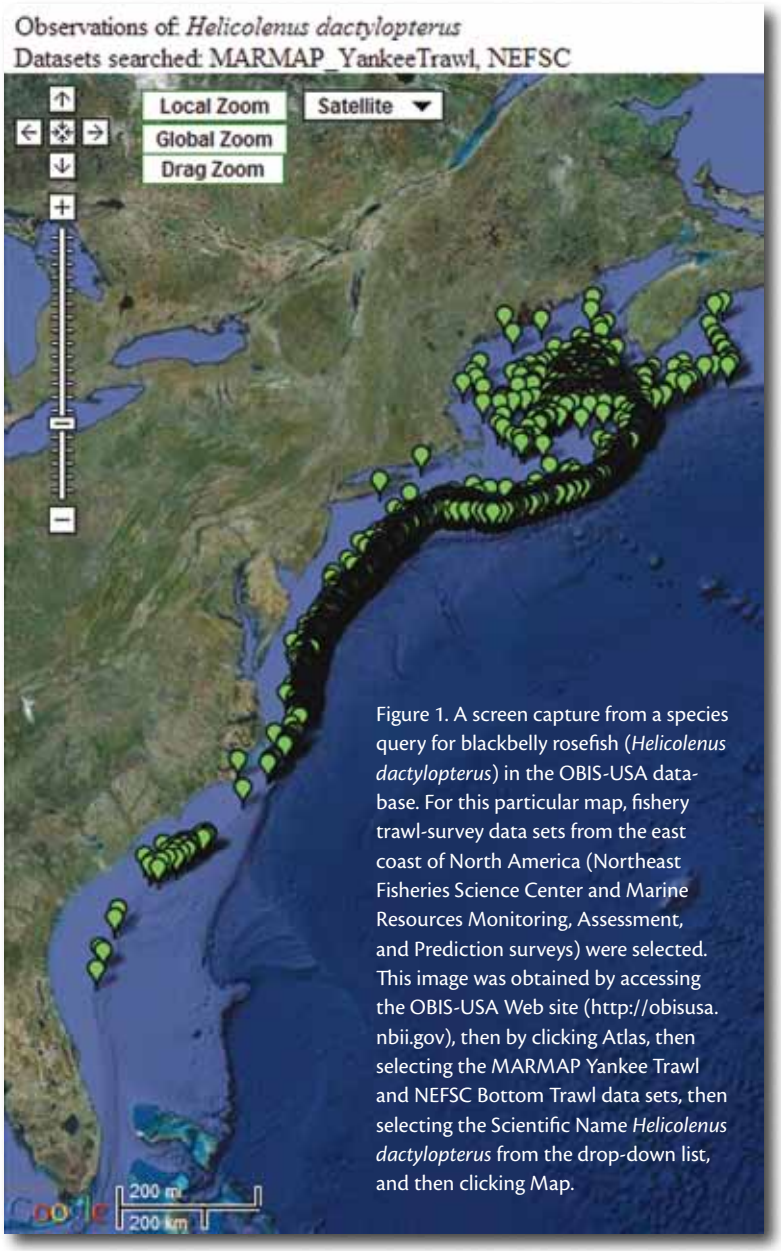


Figure 1. A screen capture from a species query for blackbelly rosefish (*Helicolenus dactylopterus*) in the OBIS-USA database. For this particular map, fishery trawl-survey data sets from the east coast of North America (Northeast Fisheries Science Center and Marine Resources Monitoring, Assessment, and Prediction surveys) were selected. This image was obtained by accessing the OBIS-USA Web site (http://obisusa.nbii.gov), then by clicking Atlas, then selecting the MARMAP Yankee Trawl and NEFSC Bottom Trawl data sets, then selecting the Scientific Name *Helicolenus dactylopterus* from the drop-down list, and then clicking Map.

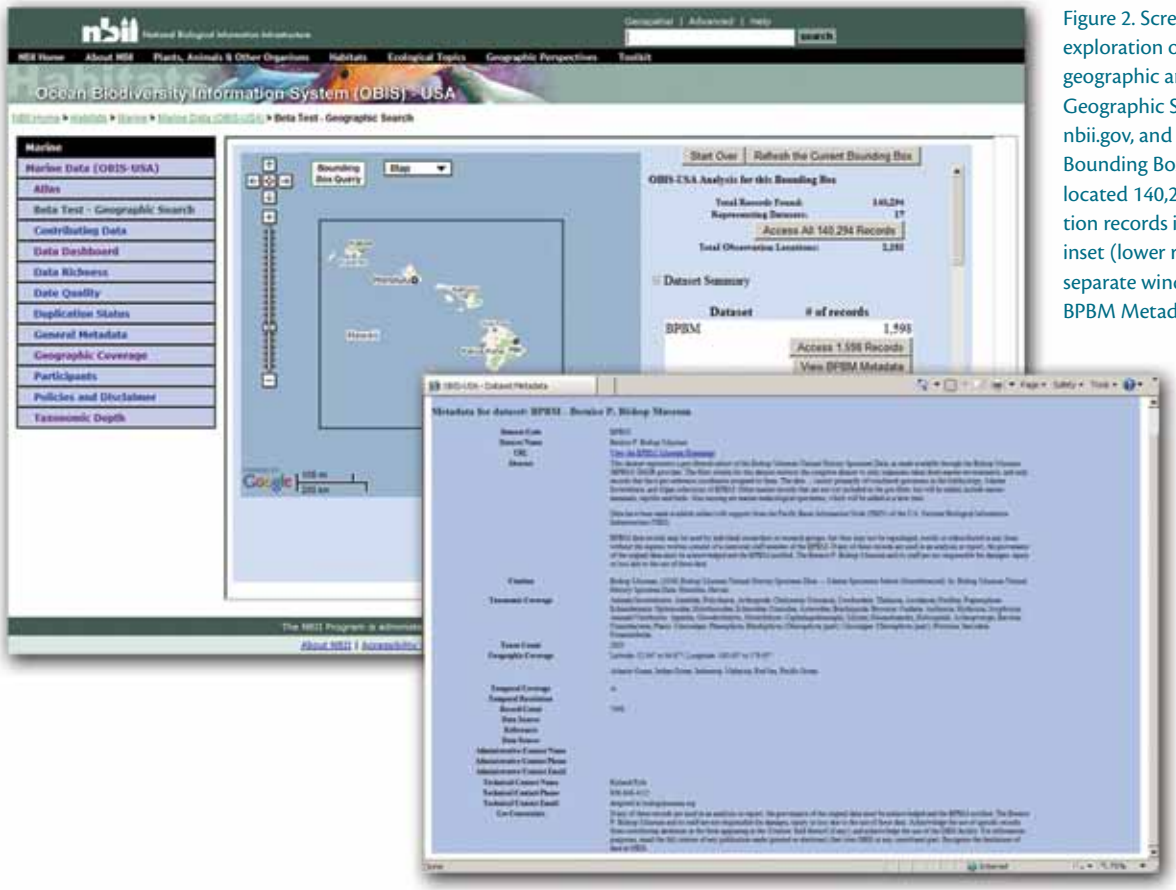


Figure 2. Screen captures from further exploration of data contained in a geographic area searched by clicking Geographic Search at http://obisusa.nbii.gov, and then conducting a Bounding Box query. The query located 140,294 species distribution records in 17 data sets. The inset (lower right, which opens as a separate window upon clicking View BPBM Metadata on the Bounding Box window) illustrates an exploration of the metadata that describe purpose, methods, and other details about the data set. In this case, metadata are shown for 1598 records from collections that were made in the area of the bounding box drawn around the main Hawaiian Islands, and that are now housed at the Bernice P. Bishop Museum (BPBM).

flexible metadata standards that meet community needs.

Ocean scientists commonly express the desire to integrate knowledge of organism distribution and demography with oceanographic and atmospheric data from multiple platforms and sources. Some fisheries surveys record environmental parameters such as temperature, salinity, dissolved oxygen, and pH with specimens sampled. But more commonly, environmental data are recorded separately from organism localities, so national or global products documenting physical properties over time must be overlain on records of organism occurrences to understand the influence of physical factors and how they vary in time and space. Easily linking an ever-growing summary of environmental layers to organism occurrence records, such as are assembled by OBIS-USA, remains a challenge. Ocean environmental data layer repositories are currently being developed by several university and agency groups (e.g., Ghent University, http://www.oracle.ugent.be; National Center for Ecological Analysis and Synthesis working group, www.nceas.ucsb.edu/projects/12504; Pacific Marine Environmental Laboratory, http://www.pmel.noaa.gov/datalinks.html). The value of environmental variables obtained from ocean observing networks and other environmental monitoring efforts must be compared with those measurements obtained synoptically with organism collections. When those environmental values agree within desired tolerances, a major objective of an ocean biogeographic information system will have been achieved.

Another challenge is to include abundance data, which will inform population and community ecology, and serve natural resources management and conservation. The OBIS schema can accommodate simple measures of abundance, but is currently insufficient for the more detailed knowledge captured by many fisheries monitoring surveys. OBIS-USA is collaborating with the NOAA Integrated Ocean Observing System, the NOAA Pacific Island Fisheries Science Center, and the University of Hawaii to lead the way in

creating interoperability between diverse biodiversity and ecological data.

## CONCLUSION

The 10-year Census of Marine Life just ended. New species were discovered and vast areas of the sea were explored, adding to knowledge of global patterns of marine biodiversity (Costello et al., 2010). As initially envisioned in *Oceanography* (Volume 13, Number 3) in 2000, OBIS and its nodes (including OBIS-USA) not only served as repositories of data from this project but also constitute the commencement of the online atlas of marine organisms.

OBIS-USA represents a national-scale effort to provide an ever-growing knowledge base that links related resources. In so doing, OBIS-USA both preserves and provides access to heretofore difficult-to-access data; it makes them interoperable through careful adherence to metadata standards and interfaces. This readily available compendium of marine biodiversity allows some of society's most complex problems to begin to be addressed by serving data in accessible formats needed by managers and policymakers (e.g., Bjorndal et al., 2011) that can be used with decision-support tools (e.g., in coastal and marine spatial planning). OBIS-USA will enhance research, as well, by pointing to gaps in knowledge, including taxa and geographical regions in need of additional exploration; such efforts are already underway using OBIS data (e.g., Mora et al., 2008).

By mobilizing data from institutions with diverse cultures and histories, OBIS-USA has begun to meet the challenge of open science and open data across a broad participant spectrum. A community and culture of data sharing, initially envisioned over 10 years ago, has been achieved for a common good and will support shared missions in monitoring and assessing ocean biological resources.

## ACKNOWLEDGEMENTS

This summary of OBIS-USA and associated programs evolved from, and was improved by, discussions with many participants in the Census of Marine Life and OBIS. In particular, we thank past and present members of the US National Committee (USNC) for the Census. Support for our work and the USNC was provided by the Alfred P. Sloan Foundation and the National Oceanic and Atmospheric Administration's Office of Ocean Exploration and Research; USNC was coordinated by the Consortium for Ocean Leadership. This work was also supported by the United States Geological Survey and the Bureau of Ocean Energy Management, Regulation and Enforcement as well as our respective agencies and universities.

## REFERENCES

Bjorndal, K.A., B.W. Bowen, M. Chaloupka, L.B. Crowder, S.S. Heppell, C.M. Jones, M.E. Lutcavage, D. Policansky, A.R. Solow, and B.E. Witherington. 2011. Better science needed for restoration in the Gulf of Mexico. *Science* 331(6017):537–538.

Collette, B.B., and M. Vecchione. 1995. Interactions between fisheries and systematics. *Fisheries* 20(1):20–25.

Constable, H., R. Guralnick, J. Wieczorek, C. Spencer, A.T. Peterson, and the VertNet Steering Committee. 2010. VertNet: A new model for biodiversity data sharing. *PloS Biology* 8(2):e1000309, doi:10.1371/journal.pbio.1000309.

Costello, M.J., M. Coll, R. Danovaro, P. Halpin, H. Ojaveer, and P. Miloslavich. 2010. A census of marine biodiversity knowledge, resources, and future challenges. *PloS ONE* 5(8):e12110, doi:10.1371/journal.pone.0012110.

Fornwall, M.D. 2000. Planning for OBIS: Examining relationships with existing national and international biodiversity information systems. *Oceanography* 13(3):31–38. Available online at: http://www.tos.org/oceanography/issues/issue_archive/vol_13.html (accessed April 19, 2011).

Grassle, J.F. 2000. The Ocean Biogeographic Information System (OBIS): An on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. *Oceanography* 13(3):5–7. Available online at: http://www.tos.org/oceanography/issues/issue_archive/vol_13.html (accessed April 19, 2011).

Mora, C., D.P. Tittensor, and R.A. Myers. 2008. The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proceedings of the Royal Society B–Biological Sciences* 275:149–155.

National Research Council. 1993. *A Biological Survey for the Nation*. Committee on the Formation of the National Biological Survey, National Academy Press, Washington, DC, 205 pp.

Tittensor, D.P., C. Mora, W. Jetz, H.K. Lotze, D. Ricard, E. Vanden Berghe, and B. Worm. 2010. Global patterns and predictors of marine biodiversity across taxa. *Nature* 466:1,098–1,103, doi:10.1038/nature09329.

Yarincik, K., and R. O'Dor. 2005. The Census of Marine Life: Goals, scope and strategy. *Scientia Marina (Barcelona)* 69(suppl. 1):201–208.